

Multi-View Kernel Consensus For Data Analysis and Signal Processing

Moshe Salhov¹, Ofir Lindenbaum², Avi Silberschatz³, Yoel Shkolnisky⁴, Amir Averbuch

¹School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

²School of Engineering, Tel Aviv University, Tel Aviv 69978, Israel

³Department of Computer Science, Yale University, New Haven, CT 06520-8285, USA

⁴School of Mathematical Science, Tel Aviv University, Tel Aviv 69978, Israel

June 29, 2016

Abstract

The input data features set for many data driven tasks is high-dimensional while the intrinsic dimension of the data is low. Data analysis methods aim to uncover the underlying low dimensional structure imposed by the low dimensional hidden parameters by utilizing distance metrics that consider the set of attributes as a single monolithic set. However, the transformation of the low dimensional phenomena into the measured high dimensional observations might distort the distance metric. This distortion can effect the desired estimated low dimensional geometric structure. In this paper, we suggest to utilize the redundancy in the attribute domain by partitioning the attributes into multiple subsets we call views. The proposed methods utilize the agreement also called consensus between different views to extract valuable geometric information that unifies multiple views about the intrinsic relationships among several different observations. This unification enhances the information that a single view or a simple concatenations of views provides.

1 Introduction

Kernel methods constitute of a wide class of algorithms for non-parametric data analysis of massive high dimensional data. Typically, a limited set of underlying factors generates high dimensional observable parameters via a non-linear mapping.

Multi-Dimensional scaling (MDS) [6, 11] has become a basis for many kernel methods. MDS is used for analyzing and visualizing data, when only pairwise similarities or dissimilarities between data points are observable. The given similarity information is depicted as pairwise distances in a low-dimensional space.

Kernel methods are based on an affinity kernel construction that encapsulates the relations (distances, similarities or correlations) among multidimensional data points. Spectral analysis of this kernel provides an efficient representation of the data that enables its analysis. The non-parametric nature of those methods enables us to uncover hidden structures in the data.

Methods such as Isomap [25], LLE [20], Laplacian eigenmaps [1], Hessian eigenmaps [8] and local tangent space alignment [27, 28] extend the MDS paradigm by considering the manifold assumption. Under this assumption, the data is assumed to be sampled from an intrinsic low dimensional manifold that captures the dependencies between observable parameters. The corresponding spectral-based embedded spaces computed by these methods identify the geometry of the manifold that incorporates the underlying factors in the data.

Similarity assessments between members in datasets is a crucial task for the analysis of any dataset. Important and popular kernel methods such as the methods discussed above utilize similarity metrics that is based on the l_2 norm between features. For example, the widely used Gaussian kernel is based on a scaled l_2 norm between multidimensional data points. In many cases, the given dataset includes redundant features that relate to the underlying factors via an unknown transformation.

Since the unknown transformation function may have zero derivatives, similarity between transformed data points might be a distorted version of the similarity between the underlying factors. Furthermore, the transformation may be done in the presence noise, which add

additional distortion to the similarity assessment. The utilization of this distorted similarity for kernel based data analysis may fail to uncover the desired geometry for the analysis.

In this paper, we propose two methods for high dimensional data analysis that aim to compensate for the distortions induced by the transformation to the similarity assessment. Both methods consider subsets of features, where each subset is defined as a view, to compute similarity assessment for each view and the agreement between all the computed similarities. Furthermore, we utilize this agreement to estimate the corresponding inaccessible pairwise similarities of the underlying factors.

Learning from several views has motivated various studies that have focused on classification and clustering that are based on the spectral characteristics of multiple datasets. Among these studies are Bilinear Model [9] and Canonical Correlation Analysis [4]. These methods are effective for clustering but neither provide a low dimensional geometry nor a structure for each view. An approach similar to Canonical Correlation Analysis [4], which seeks a linear transformation that maximizes the correlation among the views, is described in [2]. Data modeling by a bipartite graph is given in [7]. Then, based on the ‘minimum-disagreement’ algorithm, [7] partitions the dataset. Recently, a few kernel-based methods propose a model of co-regularizing kernels in both views [12]. It is done by searching for an orthogonal transformation, which maximizes the diagonal terms of the kernel matrices obtained from all views, by adding a penalty term that incorporates the disagreement among the views. A mixture of Markov chains is proposed in [29] to model the multiple views in order to apply spectral clustering. A way to incorporate given multiple metrics for the same data using a cross diffusion process is given in [26]. Again, the applicability of the suggested approach is limited only for a clustering task.

This work considers the following cases that depend on the given data:

1. The data consists of time instances that are the result by a dynamical process;
2. The covariance at each neighborhood of a multidimensional data point is accessible.

In the first case, we assume that the data is generated by a dynamical process that operates on a low dimensional manifold such as in [10, 19, 16, 15, 17] to name some. Another

interesting and relevant approach is described in [22, 21], which is a spectral approach for solving linear and non-linear Independent Component Analysis (ICA) problems. This work assumes that the data is generated by a dynamical processes in order to obtain a unique solution to a general ill-posed non-linear ICA problem. This assumption enables us to compute the local Jacobian-based distortion metric induced by a non-linear transformation that maps the parameter space into an observable space. As shown in [21], a spectral approach for solving the non-linear ICA problem can then be employed by using the Jacobian-based metric to construct a diffusion kernel. The eigenvectors of the constructed kernel, which can be chosen as proposed in [21], represent the data in terms of its independent parameters. Several applications for this approach are described in [24, 13]. The dynamical process assumption allows to isolate a common process across views. The uncommon dynamical processes are regarded as interference.

In the second case, we assume to have some knowledge about the data points that allows us to compute the covariance matrix of the local neighborhood at each data point. In this case, we utilize the existence of the relation between the Mahalanobis distances [18] in the extrinsic domain and the Mahalanobis distances in the intrinsic domain. Furthermore, we utilize this relation to define a similarity distance that considers the inherent structure in different views.

The paper has the following structure. Section 2 provides the problem formulation of the multi-view embedding. Section 3 details the multi-view analysis for dynamical processes. The analysis of a dataset with an assumed accessible covariance matrix per neighborhood is described in Section 4. Section 5 presents numerical examples. Section 6 provides concluding remarks and future work directions.

2 Problem Formulation

In the following, $\|\cdot\|$ denotes the standard Euclidean vector norm and $\|\cdot\|_F$ denotes the Frobenius matrix norm. Vectors are denoted by **Bold** letters and vector components are denoted by a superscript $[\cdot]^r$.

Let \mathcal{M} be a low-dimensional manifold that lies in the high-dimensional ambient Euclidean space \mathbb{R}^m and let $d \ll m$ be its intrinsic dimension. Let $M \subseteq \mathcal{M}$ be a dataset of $|M| = n$ multidimensional data points that were sampled from \mathcal{M} . For data analysis tasks in general and signal processing tasks in particular, each extracted/measured feature vector $\mathbf{x}_i \in M$ is assumed to have a corresponding vector $\boldsymbol{\theta}_i \in \mathbb{R}^d$ of inaccessible controlling parameters whose r th component is θ_i^r , $i = 1, \dots, n$.

Kernel methods analyze datasets such as M by exploring the geometry of the manifold \mathcal{M} whose data points were sampled as in [5, 14, 20, 25]. The computed kernel describes a measure of data points pair-wise similarity. The Euclidean norm-based similarity metric between two data points $\mathbf{x}_i, \mathbf{x}_j \in M$ is given by

$$K_\varepsilon(\mathbf{x}_i, \mathbf{x}_j) = h\left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\varepsilon}\right), \quad i, j = 1, \dots, n \quad (2.1)$$

where ε is the kernel width and $h : \mathbb{R} \rightarrow \mathbb{R}$ is a function designed such that the kernel matrix is symmetric and positive semi-definite. However, instead of using the measured features for the similarity assessment in Eq. 2.1, we want to replace the Euclidean norm $\|\mathbf{x}_i - \mathbf{x}_j\|$ with the Euclidean norm $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|$, $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j \in \mathbb{R}^d$ that corresponds directly to the similarity between the inaccessible corresponding controlling parameters instances.

In this paper, we approximate the Euclidean distance $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|$ (or the corresponding Mahalanobis distance) by utilizing the redundancy in the feature space. In order to quantify the relation between the approximated distance and the actual one, we introduce the notion of a View, which is given in Definition 2.1.

Definition 2.1 (A View). *Let I_l be a subset of m_l feature indexes that is selected from the set of m features given in dataset M . M_l is a view of M if for every $x_i \in M, 1 \leq i \leq n$ the vector $\tilde{\mathbf{x}}_{i,l} \in M_l$ is the subset of x_i that corresponds to the set of indexes in I_l .*

Under the multi-view formulation, by selecting ζ subsets of features from the features of M we generate ζ views of M , $1 \leq l \leq \zeta$. Let \mathbf{x}_i be the concatenation of the ζ views such that $\mathbf{x}_i = \cup_{l=1}^{\zeta} \tilde{\mathbf{x}}_{i,l}$ where $\tilde{\mathbf{x}}_{i,l} \triangleq [x_{i,l}^1, \dots, x_{i,l}^{m_l}]$. We further assume that the l th view is the outcome of the function $\mathbf{f}_l : \mathbb{R}^d \times \mathbb{R}^{k_l} \rightarrow \mathbb{R}^{m_l}$ such that $\tilde{\mathbf{x}}_{i,l} = \mathbf{f}_l(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i,l})$, where $\boldsymbol{\psi}_{i,l} \in \mathbb{R}^{k_l}$ is the k_l -dimensional view-specific controlling parameters.

Our goal in this work is to find a function $G : \mathbb{R}^\zeta \rightarrow [0, 1]$ such that the desired kernel similarity $K_\varepsilon(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$, $i, j = 1, \dots, n$ is approximated such that

$$K_\varepsilon(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \approx G(K_{\varepsilon_1}(\tilde{\mathbf{x}}_{i,1}, \tilde{\mathbf{x}}_{j,1}), \dots, K_{\varepsilon_l}(\tilde{\mathbf{x}}_{i,l}, \tilde{\mathbf{x}}_{j,l}), \dots, K_{\varepsilon_\zeta}(\tilde{\mathbf{x}}_{i,\zeta}, \tilde{\mathbf{x}}_{j,\zeta})) \quad (2.2)$$

where ε_l is adapted to the characteristics of the l th view and K_ε is defined in Eq. 2.1. We call this kernel the multi-view kernel.

We aim to find the underlying intrinsic geometry of \mathcal{M} . By identifying the consensus between the ζ views, we are able to approximate a kernel that is related to the inaccessible controlling parameters $\boldsymbol{\theta}_i$.

3 Multi-View of Dynamical Process

In this section, we analyze multi-views generated by a dynamical system. The following analysis extends the work of [22] and adapts the generic state-space formalism from [22, 24] to a variety of applications. We assume that the data points in M are the outputs from non-linear functions of independent stochastic Itô processes. Assume that $\boldsymbol{\theta}_i$ are samples from the Itô process $\boldsymbol{\theta}$ such that $\boldsymbol{\theta}_i \triangleq \boldsymbol{\theta}[t_i]$, $1 \leq i \leq n$, where t_i is a time instance. The dynamics of this process is described by normalized stochastic differential equations of the form

$$d\theta^r = a^r(\theta^r)dt + dw^r, \quad r = 1, \dots, d \quad (3.1)$$

where $a^r \leq \infty$ are the unknown drift coefficients and w^r are the independent white noises. For simplicity, we consider here processes that were normalized to have a unit variance noises.

The d -dimensional vector $\boldsymbol{\theta}_i$ is inaccessible. We observe its non-linear noisy mapping

$$\tilde{\mathbf{x}}_{i,l} = \mathbf{f}_l(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i,l}), \quad l = 1, \dots, \zeta, \quad i = 1, \dots, n, \quad (3.2)$$

where \mathbf{f}_l is assumed to be differentiable, bi-Lipschitz and $\boldsymbol{\psi}_{i,l} \triangleq \boldsymbol{\psi}_l[t_i]$ is a k_l -dimensional sample from an Itô process that is given by

$$d\psi_l^r = a^{d+r}(\psi_l^r)dt + dw_l^{d+r}, \quad r = 1, \dots, k_l, \quad l = 1, \dots, \zeta. \quad (3.3)$$

We consider ψ_l as an underlying interference process that corrupts the l th view. The following corollary quantify the probability of each independent process ψ_l^r to return to its starting point after time $T \leq \infty$.

Corollary 3.1. *Let $\psi_{i,l}^r$ and $\psi_{j,l}^r$ be two r th elements of the k_l -dimensional sample from $\psi_{l,i}$, where without loss of generality $i > j$ and $t_i - t_j = T$. Then, for every $\varepsilon_1 > 0$ we have $P(|\psi_{i,l}^r - \psi_{j,l}^r| \leq \varepsilon_1) > 0$.*

Proof. From the definition of the Itô process we have

$$\psi_{i,l}^r - \psi_{j,l}^r = \int_{t_j}^{t_i} a^{d+r}(\psi_l^r) dt + \int_{t_j}^{t_i} dw_l^{d+r}, \quad (3.4)$$

where the first integral on the right is a Riemann integral and the second is an Itô integral. Since a^r and T are bounded, the integral on the drift component is bounded from above as

$$B_a = \int_{t_j}^{t_i} a^{d+r}(\psi_l^r) dt \leq B_{max} < \infty. \quad (3.5)$$

The Itô integral is a random variable with $Z \triangleq \int_{t_j}^{t_i} dw_l^{d+r} \sim \mathcal{N}(0, T)$. Hence, for every $\varepsilon_1 > 0$ there exists $\varepsilon_2 > 0$ for which the density of $\{Z | |Z| \leq |B_{max}|\}$ is larger then ε_2 . Furthermore, for every instance of B_a we have

$$P(|B_a - Z| \leq \varepsilon_1) \geq \varepsilon_2^2. \quad (3.6)$$

Hence, $P(|B_a - Z| \leq \varepsilon_1) > 0$. □

In the multi-view perspective, the intrinsic controlling parameters, which are common to all the views, are considered by θ_i and the intrinsic controlling parameters, which are specific to the l th view, are denoted by $\psi_{i,l}$. We call θ_i the intrinsic parameters of the consensus for the i th sample.

We assume that the data points in M reside on several patches located on the low dimensional underlying manifold. On the other hand, if the data is spread sparsely over the manifold in the high-dimensional ambient space, then the application of an affinity kernel to the data will not reveal any patches/clusters. In this case, the data is too sparse to represent

or identify the underlying manifold structure and the only available processing tools are variations of nearest-neighbor-type algorithms. Therefore, data points on a low-dimensional manifold in a high-dimensional ambient space can either reside in locally-defined patches and then the methods in this paper are applicable to it, or scattered sparsely all over the manifold and thus there is no detectable coherent physical phenomenon that can provide an underlying explanation for it.

Given multi-view measurements, the dynamics of the consensus $\boldsymbol{\theta}$ process can be identified and its underlying geometry is revealed as described below. Each view $\tilde{\mathbf{x}}_{i,l}$ satisfies the stochastic dynamics given by the Itô Lemma for $i = 1, \dots, n$, and $l \leq \zeta$ such that

$$dx_{i,l}^r = \sum_{k=1}^d \left[\left(\frac{1}{2} \frac{\partial^2 f_l^r(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i,l})}{\partial \theta^k \partial \theta^k} + a^k \frac{\partial f_l^r(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i,l})}{\partial \theta^k} \right) dt + \frac{\partial f_l^r(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i,l})}{\partial \theta^k} dw^k \right] + \sum_{k=1}^{k_l} \left[\left(\frac{1}{2} \frac{\partial^2 f_l^r(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i,l})}{\partial \psi_l^k \partial \psi_l^k} + a^{d+k} \frac{\partial f_l^r(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i,l})}{\partial \psi_l^k} \right) dt + \frac{\partial f_l^r(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i,l})}{\partial \psi_l^k} dw^{d+k} \right]. \quad (3.7)$$

The accessible (r, k) th elements of the $m_l \times m_l$ covariance between the samples of the l th view $[C_{\tilde{\mathbf{x}}_{i,l}}]_{r,k} = \left[E_{dw} \left[(d\mathbf{x}_{i,l}) (d\mathbf{x}_{i,l})^T \right] \right]_{r,k}$, is given by

$$[C_{\tilde{\mathbf{x}}_{i,l}}]_{r,k} = \sum_{q=1}^d \frac{\partial f_l^r(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i,l})}{\partial \theta^q} \frac{\partial f_l^k(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i,l})}{\partial \theta^q} + \sum_{q=1}^{k_l} \frac{\partial f_l^r(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i,l})}{\partial \psi_l^q} \frac{\partial f_l^k(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i,l})}{\partial \psi_l^q}, \quad (3.8)$$

where $k, r = 1, \dots, m_l$. We utilize the fact that $E[dw^k dw^r] = 0$ for $k \neq r$ since the dw^r is independent of dw^k in this case. The covariance matrix can be reformulated in a matrix form as a function of the corresponding Jacobian $J_{\tilde{\mathbf{x}}_{i,l},l}$ of $f_l(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i,l})$ by

$$C_{\tilde{\mathbf{x}}_{i,l}} = J_{\tilde{\mathbf{x}}_{i,l},l} J_{\tilde{\mathbf{x}}_{i,l},l}^T, \quad i = 1, \dots, n, \quad l = 1, \dots, \zeta. \quad (3.9)$$

Lemma 3.2. *Let $\tilde{\mathbf{x}}_{i,l}$ be a noisy measurement of an Itô process according to Eq. 3.2. Let $\tilde{\mathbf{x}}_{i,l}$ and $\tilde{\mathbf{x}}_{j,l}$ be two data points from the l th view and let $\delta > 0$ be a given threshold. Furthermore, assume that the interference in the l th view $\boldsymbol{\psi}_{i,l}$ is independent of both the interference of any other view and independent of $\boldsymbol{\theta}_i$. As the number of views ζ grows, the minimal Mahalanobis distance (over the entire set of views) approaches the Euclidean distance between the corresponding governing parameters as follows:*

$$\lim_{\zeta \rightarrow \infty} \min_{1 \leq l \leq \zeta} \frac{1}{2} (\tilde{\mathbf{x}}_{i,l} - \tilde{\mathbf{x}}_{j,l})^T \Lambda_{i,j,l}^{-1} (\tilde{\mathbf{x}}_{i,l} - \tilde{\mathbf{x}}_{j,l}) \rightarrow \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|^2 + \phi, \quad (3.10)$$

where $\phi = O(\|\tilde{\mathbf{x}}_{i,l} - \tilde{\mathbf{x}}_{j,l}\|^4)$ and $\Lambda_{i,j,l}^{-1}$ is given by,

$$\Lambda_{i,j,l}^{-1} \triangleq C_{\tilde{\mathbf{x}}_{i,l}}^{-1} + C_{\tilde{\mathbf{x}}_{j,l}}^{-1}, \quad (3.11)$$

and where $C_{\tilde{\mathbf{x}}_{i,l}}$ and $C_{\tilde{\mathbf{x}}_{j,l}}$ are the covariances matrices that correspond to the l th view from data points $\tilde{\mathbf{x}}_{i,l}$ and $\tilde{\mathbf{x}}_{j,l}$, respectively.

Proof. From the assumption that \mathbf{f}_l bi-Lipschitz, \mathbf{f}_l has an inverse function $\mathbf{g}_l : \mathbb{R}^{m_l} \rightarrow \mathbb{R}^d \times \mathbb{R}^{k_l}$ such that $[\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i,l}] = \mathbf{g}_l(\tilde{\mathbf{x}}_{i,l})$. The relation between the Jacobian and the covariance matrix in Eq. 3.9 is proposed in [22] to approximate the distance between the governing parameters using a single l -th view under no-noise assumption. We reformulate this relation to consider a noise process. Expanding the function $\mathbf{g}_l(\nu_l)$ in Taylor series at the point $\nu_l = \tilde{\mathbf{x}}_{j,l}$ provides for $1 \leq r \leq d$,

$$\theta_i^r - \theta_j^r = \sum_{k=1}^{m_l} \frac{\partial g_l^r(\tilde{\mathbf{x}}_{j,l})}{\partial \nu_l^k} (\tilde{\mathbf{x}}_{i,l}^k - \tilde{\mathbf{x}}_{j,l}^k) + \sum_{k=1, q=1}^{m_l} \frac{\partial^2 g_l^r(\tilde{\mathbf{x}}_{j,l})}{\partial \nu_l^k \partial \nu_l^q} (\tilde{\mathbf{x}}_{i,l}^k - \tilde{\mathbf{x}}_{j,l}^k) (\tilde{\mathbf{x}}_{i,l}^q - \tilde{\mathbf{x}}_{j,l}^q) + \varphi, \quad (3.12)$$

where $\varphi = O(\|\tilde{\mathbf{x}}_{i,l} - \tilde{\mathbf{x}}_{j,l}\|^3)$. Similarly, the noise process elements where $1 \leq r \leq k_l$ we have

$$\begin{aligned} \psi_{i,l}^r - \psi_{j,l}^r &= \sum_{k=1}^{m_l} \frac{\partial g_l^{r+d}(\tilde{\mathbf{x}}_{j,l})}{\partial \nu_l^k} (\tilde{\mathbf{x}}_{i,l}^k - \tilde{\mathbf{x}}_{j,l}^k) \\ &+ \sum_{k=1, q=1}^{m_l} \frac{\partial^2 g_l^{r+d}(\tilde{\mathbf{x}}_{j,l})}{\partial \nu_l^k \partial \nu_l^q} (\tilde{\mathbf{x}}_{i,l}^k - \tilde{\mathbf{x}}_{j,l}^k) (\tilde{\mathbf{x}}_{i,l}^q - \tilde{\mathbf{x}}_{j,l}^q) + \varphi. \end{aligned} \quad (3.13)$$

Expanding the function $\mathbf{g}_l(\nu_l)$ in Taylor series at the point $\nu_l = \tilde{\mathbf{x}}_{i,l}$ gives for $1 \leq r \leq d$,

$$\theta_j^r - \theta_i^r = \sum_{k=1}^{m_l} \frac{\partial g_l^r(\tilde{\mathbf{x}}_{i,l})}{\partial \nu_l^k} (\tilde{\mathbf{x}}_{j,l}^k - \tilde{\mathbf{x}}_{i,l}^k) + \sum_{k=1, q=1}^{m_l} \frac{\partial^2 g_l^r(\tilde{\mathbf{x}}_{i,l})}{\partial \nu_l^k \partial \nu_l^q} (\tilde{\mathbf{x}}_{j,l}^k - \tilde{\mathbf{x}}_{i,l}^k) (\tilde{\mathbf{x}}_{j,l}^q - \tilde{\mathbf{x}}_{i,l}^q) + \varphi. \quad (3.14)$$

Similarly, the noise process elements at the point $\nu_l = \tilde{\mathbf{x}}_{i,l}$ gives for $1 \leq r \leq k_l$

$$\begin{aligned} \psi_{j,l}^r - \psi_{i,l}^r &= \sum_{k=1}^{m_l} \frac{\partial g_l^{r+d}(\tilde{\mathbf{x}}_{i,l})}{\partial \nu_l^k} (\tilde{\mathbf{x}}_{j,l}^k - \tilde{\mathbf{x}}_{i,l}^k) \\ &+ \sum_{k=1, q=1}^{m_l} \frac{\partial^2 g_l^{r+d}(\tilde{\mathbf{x}}_{i,l})}{\partial \nu_l^k \partial \nu_l^q} (\tilde{\mathbf{x}}_{j,l}^k - \tilde{\mathbf{x}}_{i,l}^k) (\tilde{\mathbf{x}}_{j,l}^q - \tilde{\mathbf{x}}_{i,l}^q) + \varphi. \end{aligned} \quad (3.15)$$

Using Eqs. 3.12 and 3.14 to compute $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|$ and by averaging between the two results to remove all the third order terms yield,

$$\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|^2 = \frac{1}{2} (\tilde{\mathbf{x}}_{i,l} - \tilde{\mathbf{x}}_{j,l})^T \left(J_{\boldsymbol{\theta}_{i,l}} J_{\boldsymbol{\theta}_{i,l}}^T + J_{\boldsymbol{\theta}_{j,l}} J_{\boldsymbol{\theta}_{j,l}}^T \right) (\tilde{\mathbf{x}}_{i,l} - \tilde{\mathbf{x}}_{j,l}) + \phi, \quad (3.16)$$

where $J_{\theta_{i,l}}$ is a $m_l \times d$ matrix that holds the partial derivative of $\mathbf{g}_l(\nu_l)$ as $[J_{\theta_{i,l}}]_{k,r} = \frac{\partial g_l^r(x_{i,l})}{\partial \nu_l^k}$, $1 \leq r \leq d$. Furthermore, using Eqs. 3.13 and 3.15 to compute $\|\boldsymbol{\psi}_{i,l} - \boldsymbol{\psi}_{j,l}\|$ and averaging between the two results such that all the third order terms are removed such that,

$$\|\boldsymbol{\psi}_{i,l} - \boldsymbol{\psi}_{j,l}\|^2 = \frac{1}{2} (\tilde{\mathbf{x}}_{i,l} - \tilde{\mathbf{x}}_{j,l})^T \left(J_{\boldsymbol{\psi}_{i,l}} J_{\boldsymbol{\psi}_{i,l}}^T + J_{\boldsymbol{\psi}_{j,l}} J_{\boldsymbol{\psi}_{j,l}}^T \right) (\tilde{\mathbf{x}}_{i,l} - \tilde{\mathbf{x}}_{j,l}) + \phi, \quad (3.17)$$

where $J_{\boldsymbol{\psi}_{i,l}}$ is a $m_l \times k_l$ matrix that holds the partial derivative of $\mathbf{g}_l(\nu_l)$ as $[J_{\boldsymbol{\psi}_{i,l}}]_{k,r} = \frac{\partial g_l^{d+r}(x_{i,l})}{\partial \nu_l^k}$, $1 \leq r \leq k_l$. Combining Eqs. 3.16 and 3.17 gives,

$$\begin{aligned} \|\boldsymbol{\psi}_{i,l} - \boldsymbol{\psi}_{j,l}\|^2 + \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|_2^2 &= \phi \\ &+ \frac{1}{2} (\tilde{\mathbf{x}}_{i,l} - \tilde{\mathbf{x}}_{j,l})^T \left(J_{\boldsymbol{\psi}_{i,l}} J_{\boldsymbol{\psi}_{i,l}}^T + J_{\boldsymbol{\psi}_{j,l}} J_{\boldsymbol{\psi}_{j,l}}^T + J_{\boldsymbol{\theta}_i} J_{\boldsymbol{\theta}_i}^T + J_{\boldsymbol{\theta}_j} J_{\boldsymbol{\theta}_j}^T \right) (\tilde{\mathbf{x}}_{i,l} - \tilde{\mathbf{x}}_{j,l}). \end{aligned} \quad (3.18)$$

From Eq. 3.10 and from the definition of J_{l,θ_i} and J_{l,ψ_i} and using the bi-Lipschitz assumption on the function \mathbf{f}_l we get

$$\frac{1}{2} (\tilde{\mathbf{x}}_{i,l} - \tilde{\mathbf{x}}_{j,l})^T \Lambda_{i,j,l}^{-1} (\tilde{\mathbf{x}}_{i,l} - \tilde{\mathbf{x}}_{j,l}) + O(\|\tilde{\mathbf{x}}_{i,l} - \tilde{\mathbf{x}}_{j,l}\|^4) = \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|^2 + \|\boldsymbol{\psi}_{i,l} - \boldsymbol{\psi}_{j,l}\|^2. \quad (3.19)$$

The noise processes $\boldsymbol{\psi}_{i,l}$ and $\boldsymbol{\psi}_{j,l}$ are independent and contribute a non-negative distortion to the desired distance.

Furthermore, from Corollary 3.1, we have $P(|\psi_{i,l}^r - \psi_{j,l}^r| \leq \varepsilon_1) > 0$ for every $1 \leq r \leq k_l$. Since $\psi_{i,l}^{r_1}$ is independent of $\psi_{i,l}^{r_2 \neq r_1}$, hence, $Pr(\|\boldsymbol{\psi}_{i,l} - \boldsymbol{\psi}_{j,l}\| < \varepsilon_3) \geq \prod_{r=1}^{k_l} P(|\psi_{i,l}^r - \psi_{j,l}^r| \leq \varepsilon_3)$ for every $\varepsilon_3 > 0$. From Corollary 3.1, there exists ε_2 such that $\prod_{r=1}^{k_l} P(|\psi_{i,l}^r - \psi_{j,l}^r| \leq \frac{\varepsilon_3}{\sqrt{k_l}}) \geq \varepsilon_2^{2k_l} > 0$.

Hence, as the number of independent views ζ grows, the probability for not finding a view where $\|\boldsymbol{\psi}_{i,l} - \boldsymbol{\psi}_{j,l}\| \leq \varepsilon_3$ diminishes as

$$\lim_{\zeta \rightarrow \infty} \prod_{1 \leq l \leq \zeta} (1 - Pr(\|\boldsymbol{\psi}_{i,l} - \boldsymbol{\psi}_{j,l}\| \leq \varepsilon_3)) \rightarrow 0, \quad (3.20)$$

since $1 - \varepsilon_2^{2k_l} < 1$ for $0 \leq l \leq \zeta$. Hence for every $\varepsilon_4 > 0$ we have $\zeta > 0$ such that,

$$Pr\left(\min_{1 \leq l \leq \zeta} \|\boldsymbol{\psi}_{i,l} - \boldsymbol{\psi}_{j,l}\| < \varepsilon_3\right) = 1 - \left(1 - \varepsilon_2^{2k_l}\right)^\zeta > 1 - \varepsilon_4, \quad (3.21)$$

and the proof completes. \square

Lemma 3.2 suggests that similarity between all the inaccessible parameter vectors $\boldsymbol{\theta}_i$, $i = 1, \dots, n$ can be approximated by the minimal Mahalanobis distance over the entire set of ζ views. The accuracy of this approximation depends on the number of available views.

Algorithm 3.1 summarizes the intrinsic similarity of the approximation procedure. The

Algorithm 3.1: Multi-view Intrinsic Similarity Approximation

Input: Data points: $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^m$ divided into ζ views and ε

Output: The approximated intrinsic kernel $\hat{K}_\varepsilon(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ and the approximated intrinsic similarity $d_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_j}$.

- 1: **for** $l = 1$ *to* ζ *and* $1 \leq i, j \leq n$ **do**
 - Compute $C_{\tilde{\mathbf{x}}_{i,l}} \quad 1 \leq i \leq n$ using Eq. 3.8
 - Compute $d_l(\tilde{\mathbf{x}}_{i,l}, \tilde{\mathbf{x}}_{j,l}) = \frac{1}{2} (\tilde{\mathbf{x}}_{i,l} - \tilde{\mathbf{x}}_{j,l})^T (C_{\tilde{\mathbf{x}}_{i,l}}^{-1} + C_{\tilde{\mathbf{x}}_{j,l}}^{-1}) (\tilde{\mathbf{x}}_{i,l} - \tilde{\mathbf{x}}_{j,l})$
 - 2: Compute $d_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_j} = G(d_l(\tilde{\mathbf{x}}_{i,l}, \tilde{\mathbf{x}}_{j,l}))$ where $G(\cdot)$ is the minimization over ζ views
 - 3: $\hat{K}_\varepsilon(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \exp\{-d_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_j}/\varepsilon\}$
-

output $\hat{K}_\varepsilon(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \approx \exp\{-\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|^2/\varepsilon\}$ of Algorithm 3.1, can be utilized to approximate any kernel of the form of Eq. 2.1 over the set of intrinsic parameters. The computational complexity of the l th covariance computation is $O(m_l^2 N + m_l^3)$, where N is the number of neighbors per data point. The computational complexity that is required for the computation of the ζ kernel matrices is $O\left(\sum_{l=1}^{\zeta} m_l^2 n^2\right)$ and in the worst case where $m_l \approx m$ we have $O(\zeta m^2 n^2)$. Algorithm 3.1 may serve as a preprocessing step to the ICA method in [21] to reduce the contribution of the undesired interference.

4 Dataset with an accessible covariance matrix

In this section, we propose a multi-view kernel K_ε and a function G from Eq. 2.2 for a dataset M with an accessible covariance matrix at each data point. The covariance matrix is computed locally. An example of such a dataset is a dataset that is generated by extracting features from a time series. The extracted features are sorted according to the relative sampling time that is neighbors for the covariance computation. However, we assume that

the given dataset enables us to compute the covariance matrix. Our goal is to approximate the kernel affinities between data points based on multiple views of the transformed intrinsic space. Following Section 2, we assume that the l th view is the outcome of an almost everywhere differentiable function $\mathbf{f}_l : \mathbb{R}^d \times \mathbb{R}^{k_l} \rightarrow \mathbb{R}^{m_l}$ such that $\tilde{\mathbf{x}}_{i,l} = \mathbf{f}_l(\boldsymbol{\theta}_i, \boldsymbol{\psi}_{i,l})$, where in this section we assume that $\boldsymbol{\theta}_i$ are not necessarily an Itô process and $k_l = 0$. Hence, a data point of each view is strictly a function of the intrinsic parameters in the consensus $\mathbf{f}_l : \mathbb{R}^d \rightarrow \mathbb{R}^{m_l}$ and $\tilde{\mathbf{x}}_{i,l} = f_l(\boldsymbol{\theta}_i)$.

The dimension m_l of the l -view is assumed to be higher than the dimension d of the intrinsic parametric space. Furthermore, the covariance matrix maximal rank is equal to the intrinsic dimension d . Therefore, we use the Moore-Penrose pseudoinverse to compute the Mahalanobis distance for $i, j = 1, \dots, n$, $l = 1, \dots, \zeta$, as

$$d_m(\tilde{\mathbf{x}}_{i,l}, \tilde{\mathbf{x}}_{j,l}) = \frac{1}{2} (\tilde{\mathbf{x}}_{i,l} - \tilde{\mathbf{x}}_{j,l})^T \left(C_{\tilde{\mathbf{x}}_{i,l}}^\dagger + C_{\tilde{\mathbf{x}}_{j,l}}^\dagger \right) (\tilde{\mathbf{x}}_{i,l} - \tilde{\mathbf{x}}_{j,l}). \quad (4.1)$$

The Mahalanobis distance enables us to compare data points in the intrinsic space by comparing data points in the ambient space as Lemma 4.1 suggests.

Lemma 4.1. *Let \mathbf{f}_l be a bi-Lipschitz transformation such that $\tilde{\mathbf{x}}_{i,l} = f_l(\boldsymbol{\theta}_i)$ and $\tilde{\mathbf{x}}_{j,l} = f_l(\boldsymbol{\theta}_j)$ two data points from M_l . Then,*

$$d_m(\tilde{\mathbf{x}}_{i,l}, \tilde{\mathbf{x}}_{j,l}) = d_m(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) + O(\phi^4), \quad (4.2)$$

where $\phi \triangleq \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|$.

Proof. The function \mathbf{f}_l is expanded into the first order Taylor near the point $\boldsymbol{\theta}_i$ such that

$$\tilde{\mathbf{x}}_{i,l} = \tilde{\mathbf{x}}_{j,l} + J_{\tilde{\mathbf{x}}_{i,l},l}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_i) + O(\phi^2) \quad (4.3)$$

where $J_{\tilde{\mathbf{x}}_{i,l},l}$ is the Jacobian of $\mathbf{f}_l(\boldsymbol{\theta}_i)$. By using Eqs. 4.3 and 4.1, we get

$$\begin{aligned} d_m(\mathbf{f}_l(\boldsymbol{\theta}_i), \mathbf{f}_l(\boldsymbol{\theta}_j)) &= \\ \frac{1}{2} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)^T J_{\tilde{\mathbf{x}}_{i,l},l}^T \left(J_{\tilde{\mathbf{x}}_{i,l},l} C_{\boldsymbol{\theta}_i} J_{\tilde{\mathbf{x}}_{i,l},l}^T \right)^\dagger J_{\tilde{\mathbf{x}}_{i,l},l} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) &+ \\ \frac{1}{2} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)^T J_{\tilde{\mathbf{x}}_{j,l},l}^T \left(J_{\tilde{\mathbf{x}}_{j,l},l} C_{\boldsymbol{\theta}_j} J_{\tilde{\mathbf{x}}_{j,l},l}^T \right)^\dagger J_{\tilde{\mathbf{x}}_{j,l},l} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) &+ O(\phi^4). \end{aligned} \quad (4.4)$$

The term $O(\phi^3)$ was canceled due to the symmetrization in Eq. 4.4. Assume that the rank of $J_{\tilde{\mathbf{x}}_{i,l}}$ is equal to the rank of $C_{\boldsymbol{\theta}_i}$ and $C_{\boldsymbol{\theta}_i}$ is a full rank. By using $m \geq d$ we get

$$d_m(\mathbf{f}_l(\boldsymbol{\theta}_i), \mathbf{f}_l(\boldsymbol{\theta}_j)) = \frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)^T C_{\boldsymbol{\theta}_i}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) + \frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)^T C_{\boldsymbol{\theta}_j}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) + O(\phi^4) \quad (4.5)$$

that can be rewritten as

$$d_m(\tilde{\mathbf{x}}_{i,l}, \tilde{\mathbf{x}}_{j,l}) = d_m(\mathbf{f}_l(\boldsymbol{\theta}_i), \mathbf{f}_l(\boldsymbol{\theta}_j)) = \frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)^T (C_{\boldsymbol{\theta}_i}^{-1} + C_{\boldsymbol{\theta}_j}^{-1})(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) + O(\phi^4). \quad (4.6)$$

□

According to Lemma 4.1, for a small distance ϕ , the Mahalanobis distance $d_m(\tilde{\mathbf{x}}_{i,l}, \tilde{\mathbf{x}}_{j,l})$ is approximately the same in each view l such that $d_m(\tilde{\mathbf{x}}_{i,l_1}, \tilde{\mathbf{x}}_{j,l_1}) \approx d_m(\tilde{\mathbf{x}}_{i,l_2}, \tilde{\mathbf{x}}_{j,l_2})$, $l_1, l_2 = 1, \dots, \zeta$.

The result proven in Lemma 4.1, is valid only if the following assumption hold

- The function f_l is bi-Lipschitz
- The rank of $J_{\tilde{\mathbf{x}}_{i,l}}$, equals to the intrinsic dimension d

If the first assumption fails at point $\tilde{\mathbf{x}}_{i,l}$, the function f_l is not bi-Lipschitz and thus for every $C > 0$ there exists $\boldsymbol{\theta}_j$ such that $|f_l(\boldsymbol{\theta}_i) - f_l(\boldsymbol{\theta}_j)| > C|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j|$. Hence, by the definition of d_m in Eq. (4.1), $d_m(\tilde{\mathbf{x}}_{i,l}, \tilde{\mathbf{x}}_{j,l}) > d_m(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$.

If the second assumption fails at point $\tilde{\mathbf{x}}_{i,l}$, the rank of $J_{\tilde{\mathbf{x}}_{i,l}}$ is strictly smaller than d , therefore, if $\|(\boldsymbol{\theta}_j - \boldsymbol{\theta}_i)\| \neq 0$ is in the null space of $J_{\tilde{\mathbf{x}}_{i,l}}$, $\|J_{\tilde{\mathbf{x}}_{i,l}}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_i)\| = 0$. Hence, $d_m(\tilde{\mathbf{x}}_{i,l}, \tilde{\mathbf{x}}_{j,l}) = 0$ although $d_m(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \neq 0$. In the following we suggest to validate the above assumption by finding the minimal distance while constraining the estimated covariances $C_{\tilde{\mathbf{x}}_{i,l}}$, $1 \leq i \leq n$, $1 \leq l \leq \zeta$ that are used for the distance estimation to have a local rank that is larger than the intrinsic dimension.

Estimating the intrinsic dimensionality d of a dataset has gained considerable importance recently. Proposed estimation methods utilize local distances and angles [3]. In the following we utilize local PCA to estimate d [23]. This method is more naturally integrated in the proposed algorithm, Algorithm 4.1.

Algorithm 4.1 approximates the intrinsic dimension d and provides an improved affinity measure that can be used for various kernel-based methods such as DM to reveal the underlined manifold. The algorithm applies validation measures in order to circumvent and identify deviation from the required assumptions.

Algorithm 4.1: Multi-view affinity measure approximation

Input: Data points: $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^m$ divided into ζ views, ε and threshold γ .

Output: Approximated $K_\varepsilon(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$.

- 1: Calculate $C_{\tilde{\mathbf{x}}_{i,l}}$ For all $1 \leq i \leq n$
 - 2: Compute $\kappa(i, l)$ as the number of singular values of $C_{\tilde{\mathbf{x}}_{i,l}}$ that are larger than γ .
 - 3: Compute κ_m as the median of $\kappa(i, l)$, $1 \leq i \leq n$, $1 \leq l \leq \zeta$
 - 4: **for** all $1 \leq i, j \leq n$, $1 \leq l \leq \zeta$ such that $\kappa(i, l) \geq \kappa_m$ and $\kappa(j, l) \geq \kappa_m$ **do**
 - 5: Calculate $d_m(\tilde{\mathbf{x}}_{i,l}, \tilde{\mathbf{x}}_{j,l})$ using Eq. 4.1 .
 - 6: Set $K_\varepsilon(\tilde{\mathbf{x}}_{i,l}, \tilde{\mathbf{x}}_{j,l}) = \exp(-d_m(\tilde{\mathbf{x}}_{i,l}, \tilde{\mathbf{x}}_{j,l}) / \varepsilon)$
 - 7: Set $K_\varepsilon(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = G(K_\varepsilon(\tilde{\mathbf{x}}_{i,1}, \tilde{\mathbf{x}}_{j,1}), \dots, K_\varepsilon(\tilde{\mathbf{x}}_{i,\zeta}, \tilde{\mathbf{x}}_{j,\zeta}))$ where $G()$ is the find maximum operation.
 - 8: **end for**
-

The first step in Algorithm 4.1 is to compute the local covariance with computational complexity of $O(Nm_l^2)$, where N is the number of neighbors. This step is done n times. In the second step the algorithm computes the numerical rank of $C_{\tilde{\mathbf{x}}_{i,l}}$, $1 \leq i \leq n$. This step has a computational complexity cost of $O(nm_l^3)$. Next, the intrinsic dimension is estimated by the median of $\kappa(j, l)$ over $1 \leq i \leq n$, $1 \leq l \leq \zeta$ with computational complexity of $O(n\zeta)$. The final step is to compute the kernel. This computation is done by first computing the Mahalanobis distance $d_m(\tilde{\mathbf{x}}_{i,l}, \tilde{\mathbf{x}}_{j,l})$, $1 \leq i \leq n$, $1 \leq l \leq \zeta$ with a total computational complexity cost of $O\left(n \sum_{l=1}^{\zeta} m_l^3\right)$. The distances are computed over data points that corresponds to covariances with sufficiently large numerical rank. The output kernel i, j th element is estimated as the maximal (using the function $G()$) kernel elements distance over the relevant views. In cases where the bi-Lipschitz condition is violated for small number of data points, the function $G()$ can include a histogram analysis to find the

maximal accumulation point. According to Lemma 4.1, this maximal accumulation point is related to the true intrinsic distance up to a small error.

5 Experimental Results

This section describes two examples that demonstrate how the multi-view approaches are used. The first example describes an embedding of a dataset that consists of several Itô processes in the presence of noise. We show that we can single out the consensus from the noisy measurements using the method in Section 3. The second example describes a dataset embedding that consists of several views with an accessible covariance matrix. In this case, we show the advantage of the multi-view-based embedding from Section 4. It is compared to the embedding of a corresponding single-view dataset that its features are a concatenation of all feature subsets from all the views.

5.1 Noisy Multi-Views with Consensus

As a numerical example for the process described in Section 3, we consider the Brownian motion $(\theta_i^1, \theta_i^2) = (w_i^1, w_i^2)$, $i = 1, \dots, 2000$ in the unit square $[0, 1] \times [0, 1]$ with a normal reflection at the boundary. The $n = 2000$ data points are sampled uniformly from the unit square as depicted in Fig. 5.1. The intrinsic parameters θ_i are measured via the set of $\zeta = 1, \dots, 7$ views and each set contains $l = 1, \dots, \zeta$ views. Each view is a function of the consensus θ_i and a function of an additional view-specific Itô process given by $\psi_{i,l}$ that is considered as interference.

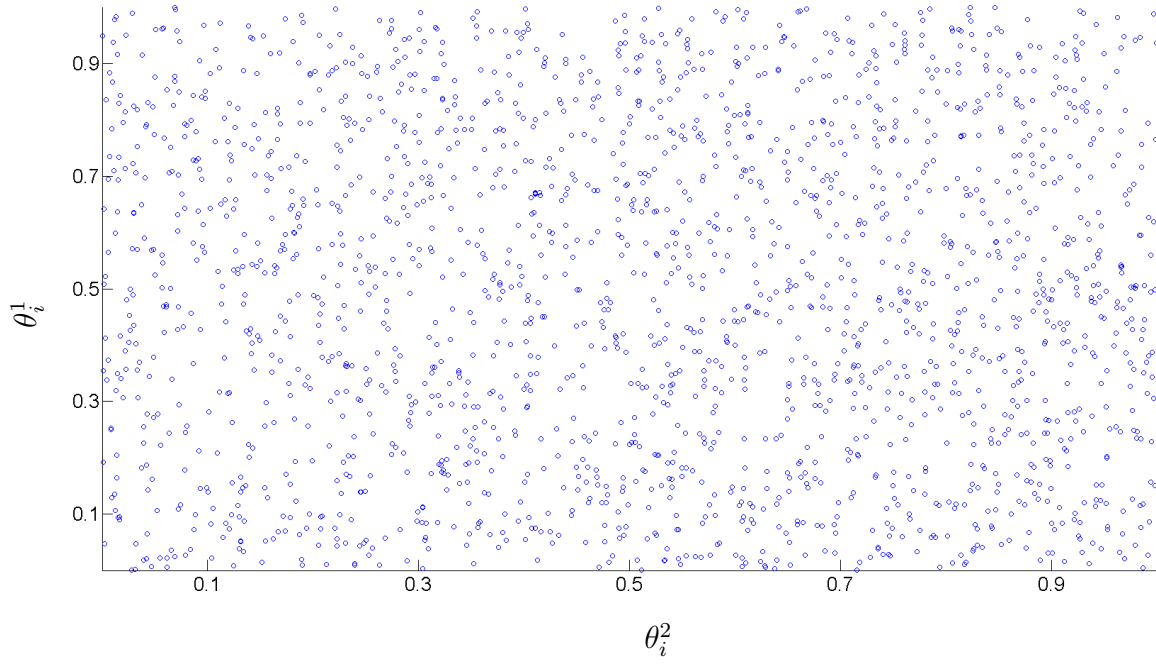


Figure 5.1: The intrinsic parameters $\boldsymbol{\theta}_i$, $i = 1, \dots, n$ of the data.

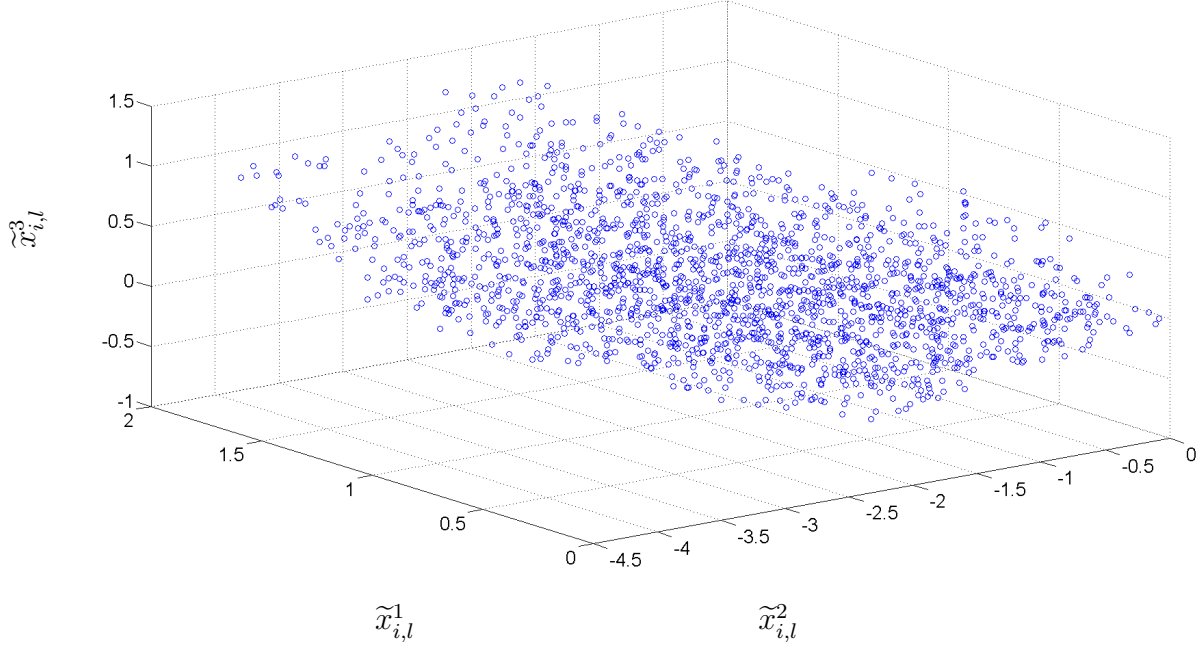


Figure 5.2: A transformed θ_i from a specific view $\tilde{\mathbf{x}}_{i,l}$.

We measured each view $\tilde{\mathbf{x}}_{i,l}$, $i = 1, \dots, n$, through a non-linear transformation by,

$$\tilde{x}_{i,l}^1 = a_{1,1,l}(\theta_i^1)^{b_{1,1,l}} + a_{1,2,l}(\theta_i^2)^{b_{1,2,l}} + a_{1,3,l}(\psi_{i,l})^{b_{1,3,l}} \quad (5.1)$$

$$\tilde{x}_{i,l}^2 = a_{2,1,l}(\theta_i^1)^{b_{2,1,l}} + a_{2,2,l}(\theta_i^2)^{b_{2,2,l}} + a_{2,3,l}(\psi_{i,l})^{b_{2,3,l}} \quad (5.2)$$

$$\tilde{x}_{i,l}^3 = a_{3,1,l}(\theta_i^1)^{b_{3,1,l}} + a_{3,2,l}(\theta_i^2)^{b_{3,2,l}} + a_{3,3,l}(\psi_{i,l})^{b_{3,3,l}}, \quad (5.3)$$

where $l = 1, \dots, \zeta$ and $a_{k,q,l}$ is random number chosen uniformly from the interval $[-2, 2]$, $k, q = 1, \dots, 3$. Furthermore, $b_{k,q,l}$ is a random integer chosen uniformly from the set $\{-3, -2, -1, 1, 2, 3\}$. We call $\tilde{\mathbf{x}}_i = (\tilde{x}_{i,l}^1, \tilde{x}_{i,l}^2, \tilde{x}_{i,l}^3)$ the observable sample from the space of the l th view. This sample depends on $\theta_i = (\theta_i^1, \theta_i^2, \psi_{i,l})$. An illustration of $\tilde{\mathbf{x}}_{i,l}$ for a specific view is given in Fig. 5.2. The specific instances of the parameters of the transformation that

are used in Fig. 5.2 are

$$\tilde{x}_{i,l}^1 = -1.94(\theta_i^1)^1 + 0.24(\theta_i^2)^{-2} - 0.62(\psi_{i,l})^1 \quad (5.4)$$

$$\tilde{x}_{i,l}^2 = -1.59(\theta_i^1)^1 + 1.39(\theta_i^2)^1 + 0.53(\psi_{i,l})^3 \quad (5.5)$$

$$\tilde{x}_{i,l}^3 = -0.68(\theta_i^1)^{-3} + 0.34(\theta_i^2)^2 + 1.10(\psi_{i,l})^{-2}. \quad (5.6)$$

Algorithm 3.1 was applied to estimate the pairwise intrinsic distances between data points of the consensus. Initially, we run $N_c = 20000$ stochastic simulations for a short time period $dt = 0.005$ that were initiated at point θ_i . The result of the N_c stochastic simulations is a point-cloud in the neighborhood of θ_i . Each point-cloud is mapped using the transformation in Eq. 5.1 to a point-cloud in the neighborhood of $\tilde{\mathbf{x}}_{i,l}$ in the measured space. Following Eq. 3.9, we calculate the 3×3 sample covariance matrix for the point-cloud of each data point in each view. Then, Eq. 3.19 is used to estimate the pairwise distances between intrinsic points in the consensus and the interference. The minimal distance (over the views) is used to estimate the pairwise distance between intrinsic points of the consensus θ_i and θ_j $i, j = 1, \dots, n$, as

$$\hat{K}_\varepsilon(\theta_i, \theta_j) \approx \min_{1 \leq l \leq \zeta} \left\{ \exp \frac{-\|\theta_i - \theta_j\|^2 - \|\psi_{i,l} - \psi_{j,l}\|^2}{\varepsilon} \right\}, \quad (5.7)$$

where $\varepsilon = 0.02$. This construction is given in Eq. 3.19 and it fits the conditions of the described experiment. The more views we have, the accuracy of the estimated pair-wise affinities improves. Let $K_\varepsilon(\theta_i, \theta_j) = \exp \frac{-\|\theta_i - \theta_j\|^2}{2\varepsilon}$ be the inaccessible and anisotropic ground truth kernel. The quality factor Q of the K_ε approximation of \hat{K}_ε is measured by the relative error in Frobenius norm

$$Q = \frac{\|K_\varepsilon - \hat{K}_\varepsilon\|_F}{\|\hat{K}_\varepsilon\|_F}. \quad (5.8)$$

Figure 5.3 displays the relation between the number of views and the approximation quality factor Q of the kernel estimation.

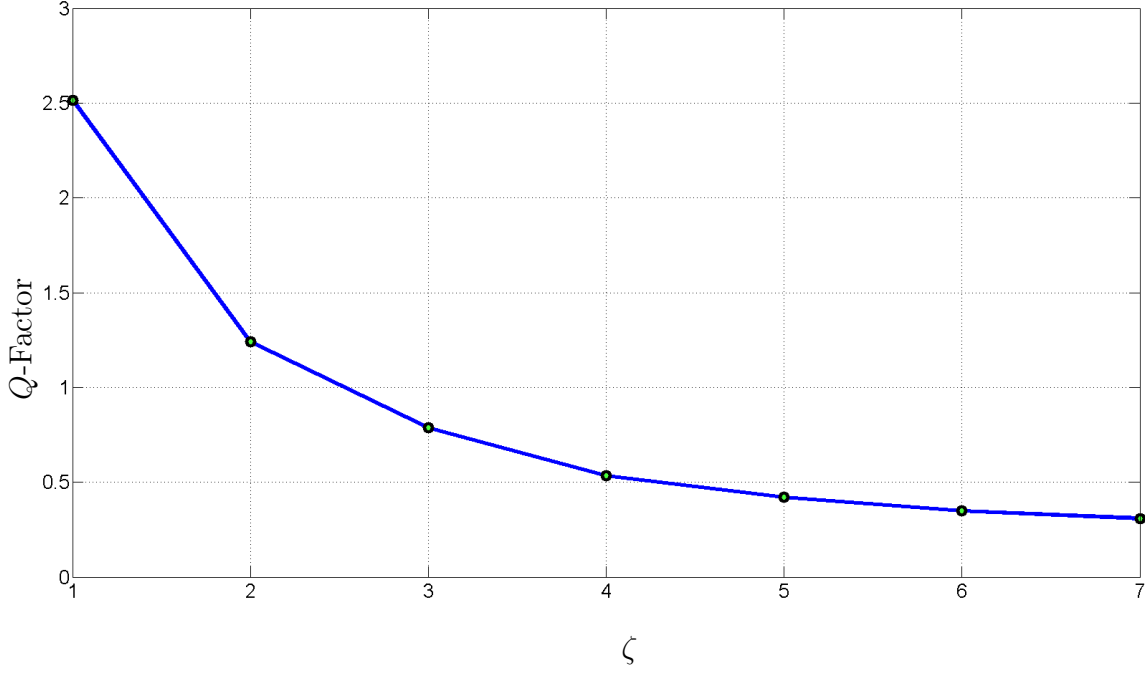


Figure 5.3: Q-factor as a function of ζ . For each value of ζ , the average accuracy is computed over 50 repetitions of the simulation.

According to [22], the anisotropic kernel K_ε converges to the Fokker-Planck operator. Hence, $D^{-1}K_\varepsilon \approx -\frac{\varepsilon}{2}L$, where D is a diagonal matrix with $[D]_{ii} = \sum_{j=1}^n K_\varepsilon(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ and L is the Fokker-Planck operator of the parametric manifold $\boldsymbol{\theta}_i^1$ and $\boldsymbol{\theta}_i^2$ in Fig. 5.1. From the definition of $\boldsymbol{\theta}_i$ in this example, the parametric manifold is the unit square with a uniform density, so that L is the Laplacian of the unit square whose eigenvalues (for the Neumann boundary conditions) are $\mu_{n,m} = \pi^2(n^2 + m^2)$, $n, m = 0, 1, \dots$. Therefore, the eigenvalues λ_i of the row stochastic matrix $D^{-1}K_\varepsilon$ are given by $\lambda_i \approx e^{-\frac{\varepsilon}{2}\mu_{n,m}}$. Figure 5.4 shows the values of $-2 \log(\lambda_i) / (\pi^2 \varepsilon)$, $i = 1, \dots, 10$, for the approximated $\hat{D}^{-1}\hat{K}_\varepsilon$, where \hat{D} is a diagonal matrix and $[\hat{D}]_{ii} = \sum_{j=1}^n \hat{K}_\varepsilon(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$. The approximation of the spectral lines $n^2 + m^2 = [0, 1, 1, 2, 4, 4, 5, 5, 8, 9]$ are shown in the Fig 5.4 where the approximated λ_i $i = 1, \dots, 10$, were computed using the multi-view scheme of Algorithm 3.1 for different

$l = 7$ noisy views. Comparing the approximated $n^2 + m^2$ with the expected ones, Fig 5.4 shows that although the parametric manifold was sampled in the presence of noise and an unknown non-linear transformation, for $i < 9$ the approximation is done with relatively small error. However, for $i > 9$ the error is larger and more views maybe needed to reduce the error.

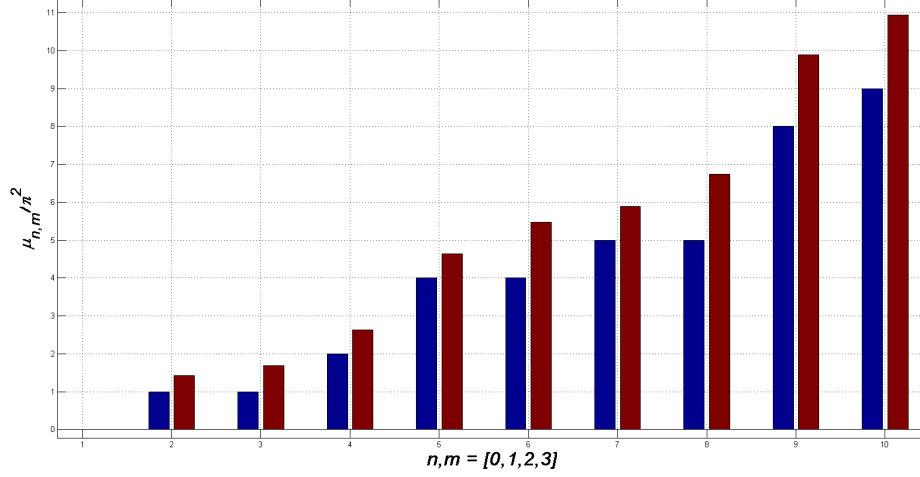


Figure 5.4: Recovery of the Fokker-Planck eigenvalues on the unit square in the presence of noisy multi-view process: theoretical spectral lines (blue) and the approximated spectral lines (red)

5.2 Multi-views with an accessible covariance matrix

Prior to the demonstration of the performance of Algorithm 4.1, we evaluate the approximation of the Mahalanobis distance in the inaccessible feature space using Eq. 4.2. First, we generate a 3-dimensional manifold using the following equation

$$\text{Helix} : \tilde{\mathbf{x}}_{i,1} = \begin{bmatrix} \tilde{x}_{i,1}^1 \\ \tilde{x}_{i,1}^2 \\ \tilde{x}_{i,1}^3 \end{bmatrix} = \begin{bmatrix} f_1^1(\theta_i) \\ f_1^2(\theta_i) \\ f_1^3(\theta_i) \end{bmatrix} = \begin{bmatrix} (2 + \cos(8\theta_i)) \cdot \cos(\theta_i) \\ (2 + \cos(8\theta_i)) \cdot \sin(\theta_i) \\ (3\theta_i^2 - \theta_i) \end{bmatrix}, \quad (5.9)$$

where the intrinsic governing parameters θ_i , $i = 1, \dots, n = 20,000$, is generated by θ_i that are uniformly drawn from $[0, 2\pi]$. The generated manifold is presented in Fig. 5.5.

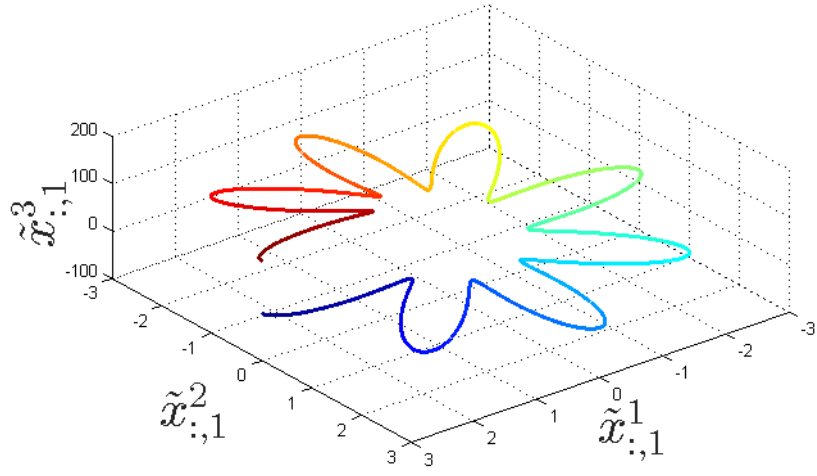


Figure 5.5: The sampled helix (Eq. 5.9) used by the experiment

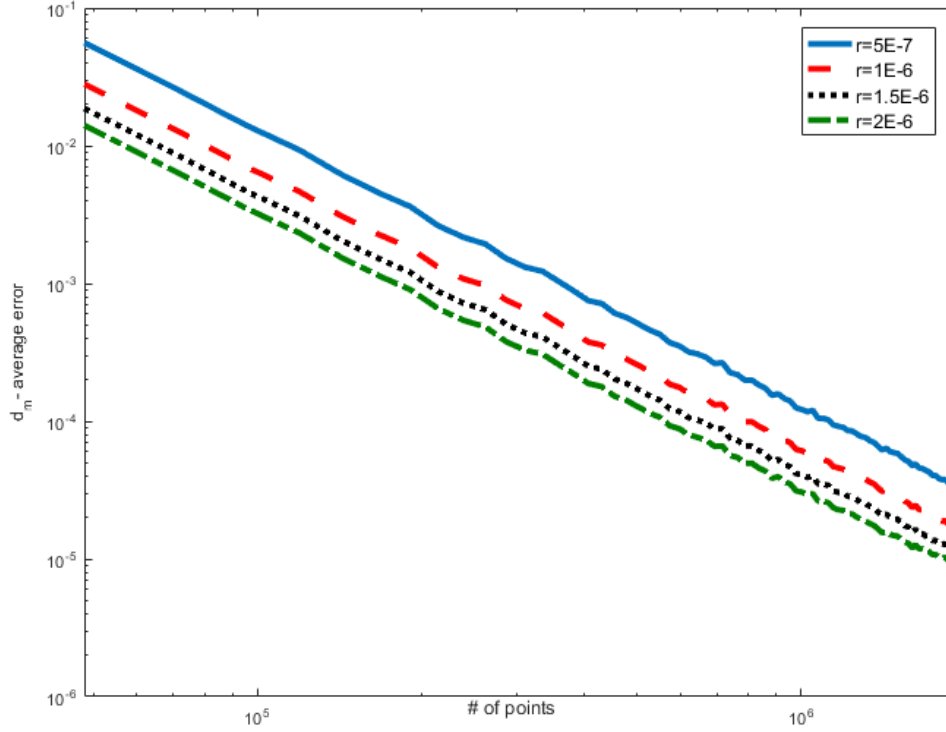


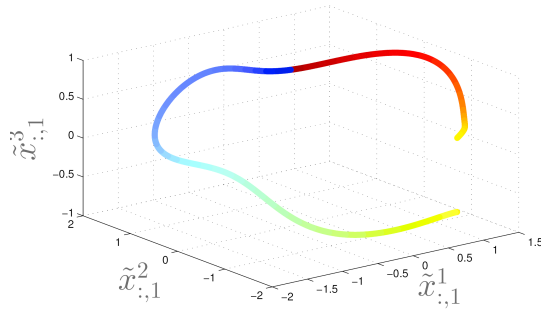
Figure 5.6: The average error between the Mahalanobis distances in the ambient space and the corresponding Mahalanobis distances in the parametric space as a function of the number of points per unit length generated in the parametric space. The value r , represents the neighborhood for computation of the covariance matrix.

The empirical covariance matrix in Eq. 4.1 depends on the neighbors of each data point. Denote by $\varepsilon_r > 0$ the neighborhood radius used by the covariance matrix computation at each point. The average error of the approximation in Eq. 4.1 is presented in Fig. 5.6, where the average is taken over 100 repeated simulations. It is evident that the errors for both datasets diminish as ε_r increases and the number of neighbors increase. As a result, the covariance matrix approximation improves and with it the similarity between both Mahalanobis distances in the ambient space and the corresponding Mahalanobis distances in the parametric space.

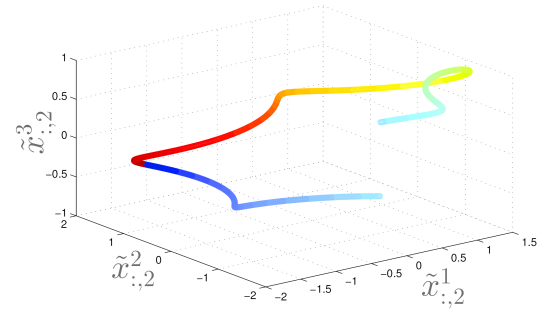
In the rest of this section, the approximation of the affinity measure is evaluated. $\zeta = 10$ views are generated to evaluate the performance of Algorithm 4.1. All the views are 3-dimensional that are based on one underlying angular parameter denoted by $\theta_i \in \mathbb{R}$, $i = 1, \dots, n$. The 10 views are generated by the application of the following function

$$\text{Helix III: } \tilde{\mathbf{x}}_{i,l} = \begin{bmatrix} \tilde{x}_{i,l}^1 \\ \tilde{x}_{i,l}^2 \\ \tilde{x}_{i,l}^3 \end{bmatrix} = \begin{bmatrix} f_l^1(\boldsymbol{\theta}_i) \\ f_l^2(\boldsymbol{\theta}_i) \\ f_l^3(\boldsymbol{\theta}_i) \end{bmatrix} = \begin{bmatrix} (4/3) \cdot \cos(\boldsymbol{\theta}_i + Z_l^1) - (1/3) \cdot \cos(4(\boldsymbol{\theta}_i + Z_l^1)) \\ (4/3) \cdot \sin(\boldsymbol{\theta}_i + Z_l^2) - (1/3) \cdot \sin(4(\boldsymbol{\theta}_i + Z_l^2)) \\ \sin(0.8 \cdot \text{mod}((\boldsymbol{\theta}_i + Z_l^3), 2\pi)) \end{bmatrix}, \quad (5.10)$$

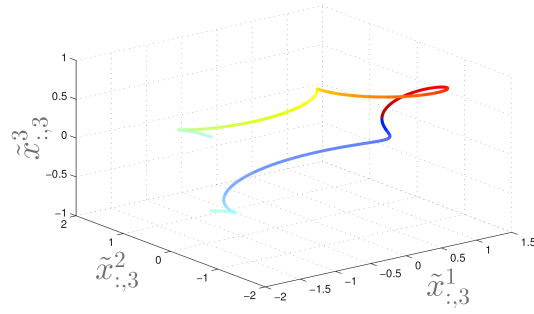
where $Z_l^1, Z_l^2, Z_l^3, 1 \leq l \leq 10$, are random variables drawn from a uniform distribution on the interval $[0, 2\pi]$. As demonstrated in Fig. 5.7, each view is a deformation of an open flower-shaped manifold. A similar deformation can occur in various real-life applications where the measured data is the output from some non-linear phenomena. In this experiment, we demonstrate the ability of a multi-view approach to overcome such deformations.



(a) View I



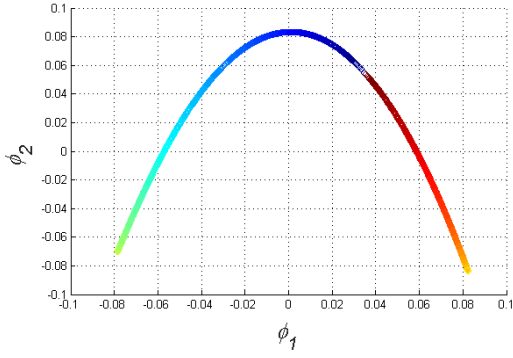
(b) View II



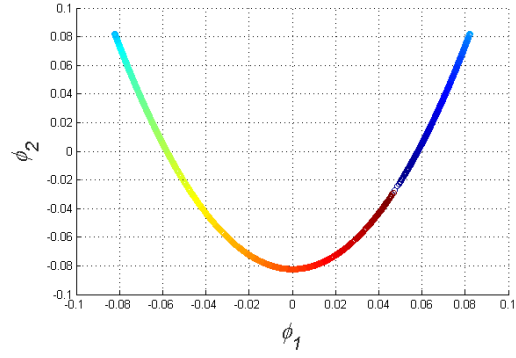
(c) View III

Figure 5.7: The three manifolds generated from the use of Eq. 5.10.

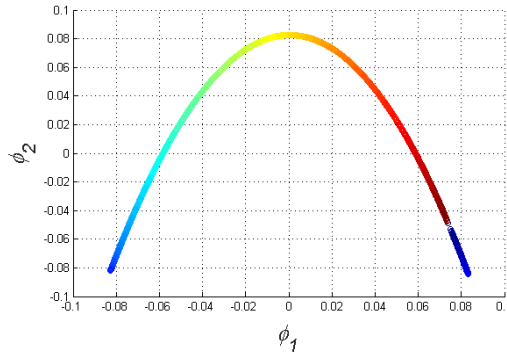
To extract the underlying parameter θ_i , we first compute the DM for each view. The two leading coordinates of the extracted embedding are denoted by ϕ_1 and ϕ_2 . They are presented in Fig. 5.8. All the extracted manifolds are horseshoe-shapes that consist of a large gap created by the deformation (Eq. 5.10) in the third coordinate of the sampled data.



(a) DM of View I



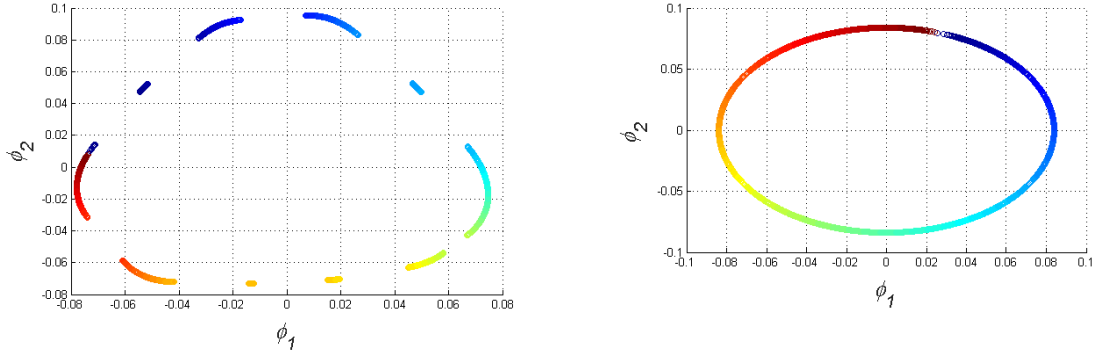
(b) DM of View II



(c) DM of View III

Figure 5.8: The DM embedding of each single view $l = 1, \dots, 3$.

Next, the 10 available views are concatenated to a single view $\mathbf{x}_i = [\tilde{\mathbf{x}}_{i,1}, \dots, \tilde{\mathbf{x}}_{i,\zeta}]$, $i = 1, \dots, n$. The Mahalanobis distance is computed using Eq. 4.1 and DM is applied to the resulted kernel. The first two leading coordinates of the kernel-based DM are presented in Fig. 5.8. The large gap in each of the extracted manifolds is a deformation caused by the third coordinate of the transformation functions \mathbf{f}_l (Eq. 5.10). Furthermore, the output from the application of DM to the concatenation of the 10 views is presented in Fig. 5.9(a). The embedded manifold is even more distorted as the gaps in the embedding suggests. Hence, the standard procedure concatenating the given set of features without considering the specific distortion each subset of features may introduce into the embedding might amplify the distortion in the resulted embedding.



(a) DM coordinates of the concatenation of views (b) DM coordinates of the multi-view-based kernel

Figure 5.9: DM coordinates of concatenation \mathbf{x}_i $i = 1, \dots, 20,000$ compared to the DM applied to the multi-view kernel, approximated by Algorithm 4.1 that considers the views $\tilde{\mathbf{x}}_{i,l}$, $l = 1, \dots, 10$.

Finally, we apply Algorithm 4.1 to all 10 views and compute the two leading DM coordinates. The outputs are presented in Fig. 5.9(b). The algorithm overcomes the deformations from all the gaps by considering only the non-deformed small distances as the outcome of the majority vote function in Algorithm 4.1. The result is the circle shaped manifold that completely agrees with the corresponding intrinsic controlling angle parameter θ_i .

6 Conclusions

This paper presents a kernel construction scheme that designs kernels by approximating the similarity between intrinsic parameters that are common to multiple subset of features (also called views) in the presence of noise and non-linear transformation. The presented method utilizes the relation between the Jacobian of each view and the corresponding Mahalanobis distance when a local covariance can be approximated. This relation enables the approximation of the affinities between intrinsic controlling parameters by considering the Mahalanobis distance of each view. The constructed kernel can further be normalized and decomposed to find an embedding of the data. In order to demonstrate the effectiveness of the pro-

posed method, we analyzed several synthetic datasets. This analysis showed that the correct affinities can be approximated in the presence of significant noise and a unknown non-linear transformation. Furthermore, in cases where the features are the output of a transformed intrinsic parameters with an associated non-full rank Jacobian then the concatenation of the entire set of features results in a deformed manifold. In this case the proposed multi-view scheme overcomes the problematic Jacobian and outputs a non-deformed manifold. The proposed methodology involves a single spectral decomposition while increasing the number of (smaller Covariance-based) Mahalanobis distances computations per affinity. Hence, the growth in computation complexity is negligible.

Acknowledgment

This research was partially supported by the Israeli Ministry of Science & Technology (Grants No. 3-9096, 3-10898), US-Israel Binational Science Foundation (BSF 2012282), Blavatnik Computer Science Research Fund and Blavatnik ICRC Funds. We also would like to thank Aviv Rotbart for assisting with the experimental results.

References

- [1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.
- [2] B. Boots and G. Gordon. Two-manifold problems with applications to nonlinear system identification. *In Proc. 29th Intl. Conf. on Machine Learning (ICML)*, 2012.
- [3] Claudio Ceruti, Simone Bassis, Alessandro Rozza, Gabriele Lombardi, Elena Casiraghi, and Paola Campadelli. Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration. *Pattern Recognition*, 47(8):2569 – 2581, 2014.

- [4] K. Chadhuri, S. M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. *In Proc. 26th Intl. Conf. on Machine Learning (ICML)*, pages 126–136, 2009.
- [5] R.R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.
- [6] T. Cox and M. Cox. *Multidimensional Scaling*. Chapman and Hall, London, UK, 1994.
- [7] V. R. de Sa. Spectral clustering with two views. *In ICML Workshop on Learning with Multiple Views*, 2005.
- [8] D.L. Donoho and C. Grimes. Hessian eigenmaps: New locally linear embedding techniques for high dimensional data. *Proceedings of the National Academy of Sciences of the United States of America*, 100:5591–5596, May 2003.
- [9] W. T. Freeman and J. B. Tenenbaum. Learning bilinear models for two-factor problems in vision. *In Computer Vision and Pattern Recognition (CVPR)*, volume 21, pages 554–560, 1997.
- [10] I. G. Kevrekidis, C. W. Gear, and G. Hummer. Equation-free: The computer-aided analysis of complex multiscale systems. *AIChE J.*, 50(7):1346–1355, 2004.
- [11] J.B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.
- [12] A. Kumar, P. Rai, and H. Daume. Co-regularized multi-view spectral clustering. *In Advances in Neural Information Processing Systems*, volume 24, 2011.
- [13] D. Kushnir, A. Haddad, and R. R. Coifman. Anisotropic diffusion on sub-manifolds with application to earth structure classification. *Applied and Computational Harmonic Analysis*, 32(2):280 – 294, 2012.
- [14] S. Lafon. *Diffusion Maps and Geometric Harmonics*. PhD thesis, Yale University, May 2004.

- [15] R. T. Li, T.-P. Tian, and S. Sclaroff. Simultaneous learning of nonlinear manifold and dynamical models for high-dimensional time series. In *ICCV*, pages 1–8. IEEE, 2007.
- [16] Ruei-Sung Lin, Che-Bin Liu, Ming-Hsuan Yang, Narendra Ahuja, and Stephen Levinson. Learning nonlinear manifolds from time series. In A. Leonardis, H. Bischof, and A. Pinz, editors, *Computer Vision ECCV 2006*, volume 3952 of *Lecture Notes in Computer Science*, pages 245–256. Springer Berlin Heidelberg, 2006.
- [17] J. H. Mack, L. Buesing, J. P. Cunningham, B. M. Yu, V. S. Krishna, and M. Sahani. Empirical models of spiking in neural populations. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1350–1358. Curran Associates, Inc., 2011.
- [18] P. C. Mahalanobis. On the generalised distance in statistics. In *Proceedings National Institute of Science, India*, volume 2, pages 49–55, April 1936.
- [19] A. Rahimi, B. Recht, and T. Darrell. Learning appearance manifolds from video. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 868–875, 2005.
- [20] S. T. Roweis and K. S. Lawrence. Nonlinear dimensionality reduction by local linear embedding. *Science*, 290:2323–2326, 2000.
- [21] A. Singer. Spectral independent component analysis. *Applied and Computational Harmonic Analysis*, 21(1):135–144, 2006.
- [22] A. Singer and R.R. Coifman. Non-linear independent component analysis with diffusion maps. *Applied and Computational Harmonic Analysis*, 25(2):226–239, 2008.
- [23] A. Singer and H.-t. Wu. Orientability and diffusion maps. *Applied and Computational Harmonic Analysis*, 31(1):44–58, 2011.

- [24] R. Talmon and R. R. Coifman. Intrinsic modeling of stochastic dynamical systems using empirical geometry. *Applied and Computational Harmonic Analysis*, 39(1):138–160, 2015.
- [25] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
- [26] B. Wang, J. Jiang, W. Wang, Z. Zhou, and Z. Tu. Unsupervised metric fusion by cross diffusion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2997–3004, 2012.
- [27] G. Yang, X. Xu, and J. Zhang. Manifold alignment via local tangent space alignment. *International Conference on Computer Science and Software Engineering*, December 2008.
- [28] Z. Zhang and H. Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *Technical Report CSE-02-019, Department of Computer Science and Engineering, Pennsylvania State University*, 2002.
- [29] D. Zhou and C. J. C. Burges. Spectral clustering and transductive learning with multiple views. *Proceedings of the 24th international conference on Machine learning*, pages 1159–1166, 2007.